



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

An alternative strategy for trypanosome survival in the mammalian bloodstream revealed through genome and transcriptome analysis of the ubiquitous bovine parasite *Trypanosoma (Megatrypanum) theileri*

Citation for published version:

Kelly, S, Ivens, A, Mott, G, O'Neill, E, Emms, D, Macleod, O, Voorheis, P, Tyler, K, Clark, M, Matthews, J, Matthews, K & Carrington, M 2017, 'An alternative strategy for trypanosome survival in the mammalian bloodstream revealed through genome and transcriptome analysis of the ubiquitous bovine parasite *Trypanosoma (Megatrypanum) theileri*' *Genome Biology and Evolution*, vol. 9, no. 8, pp. 2093-2109. DOI: 10.1093/gbe/evx152

Digital Object Identifier (DOI):

[10.1093/gbe/evx152](https://doi.org/10.1093/gbe/evx152)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

Genome Biology and Evolution

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



An alternative strategy for trypanosome survival in the mammalian bloodstream revealed through genome and transcriptome analysis of the ubiquitous bovine parasite *Trypanosoma (Megatrypanum) theileri*.

Steven Kelly^{1a}, Alasdair Ivens^{1b}, G. Adam Mott^{2b}, Ellis O'Neill^a, David Emms^a, Olivia Macleod^c, Paul Voorheis^d, Kevin Tyler^e, Matthew Clark^f, Jacqueline Matthews^g, Keith Matthews^{b*} and Mark Carrington^{c*}

^a Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK

^b Centre for Immunity, Infection and Evolution and Institute for Immunology and Infection Research, School of Biological Sciences, University of Edinburgh, Charlotte Auerbach Road, Edinburgh EH9 3FL, United Kingdom.

^c Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, United Kingdom.

^d School of Biochemistry and Immunology, Trinity College, Dublin, Ireland.

^e Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, Norfolk, NR4 7TJ, United Kingdom

^f Earlham Institute, Norwich Research Park, Norwich NR4 7UH, UK.

^g Moredun Research Institute, Pentlands Science Park, Bush Loan, Penicuik, Midlothian EH26 0PZ, United Kingdom.

¹ Contributed equally to this work

² Present address: Department of Cell & Systems Biology, University of Toronto, 25 Willcocks St., Toronto, Ontario, Canada

* Joint corresponding authors:

Keith Matthews, tel: +44 131 651 3639, email: keith.matthews@ed.ac.uk.

Mark Carrington, tel: +44 1223 333 683, email: mc115@cam.ac.uk

Abstract

There are hundreds of *Trypanosoma* species that live in the blood and tissue spaces of their vertebrate hosts. The vast majority of these do not have the ornate system of antigenic variation that has evolved in the small number of African trypanosome species, but can still maintain long term infections in the face of the vertebrate adaptive immune system. *Trypanosoma theileri* is a typical example, it has a restricted host range of cattle and other Bovinae and is only occasionally reported to cause patent disease although no systematic survey of the effect of infection on agricultural productivity has been performed. Here, a detailed genome sequence and a transcriptome analysis of gene expression in bloodstream form *T. theileri* have been performed. Analysis of the genome sequence and expression showed that *T. theileri* has a typical kinetoplastid genome structure and allowed a prediction that it is capable of meiotic exchange, gene silencing via RNA interference and, potentially, density-dependent growth control. In particular, the transcriptome analysis has allowed a comparison of two distinct trypanosome cell surfaces, *T. brucei* and *T. theileri*, that have each evolved to enable the maintenance of a long-term extracellular infection in cattle. The *T. theileri* cell surface can be modelled to contain a mixture of proteins encoded by four novel large and divergent gene families and by members of a major surface protease gene family. This surface composition is distinct from the uniform variant surface glycoprotein coat on African trypanosomes providing an insight into a second mechanism used by trypanosome species that proliferate in an extracellular milieu in vertebrate hosts to avoid the adaptive immune response.

Keywords

Trypanosoma theileri; genome; transcriptome; cell surface components

Data deposition

The Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession NBCO000000000. The version described in this paper is version NBCO010000000. Raw RNASeq reads are available from EBI ArrayExpress under the accession number E-MTAB-5327 mRNA abundances in *T. brucei* were derived from EBI ArrayExpress E-MTAB-3335 and E-MTAB-5460. The complete orthogroups dataset including all alignments and phylogenetic trees is available from the Zenodo research data repository at <https://doi.org/10.5281/zenodo.193020>

Introduction

Trypanosomatid parasites have been extensively studied over the last 100 years since their discovery as the agents of a number of important diseases of the tropics (Azizi, et al. 2016; Cox 2016; Steverding 2008, 2014). Major pathogens in this group include African trypanosomes, such as *Trypanosoma brucei*, *T. congolense* and *T. vivax*, the causal agents of Human and Animal African Trypanosomiasis; *T. cruzi* causing Human American Trypanosomiasis or Chagas' disease, and *Leishmania* species responsible for cutaneous and visceral Leishmaniasis. The importance of these arthropod borne protozoal infections has driven detailed molecular analyses of their immune evasion mechanisms, gene function, biochemistry and genome structure to the extent that they now represent the best characterised eukaryotic microbes beyond the established unicellular genetic models such as yeasts and *Dictyostelium*. In addition, the distant evolutionary divergence of this group from other eukaryotes is evidenced by the array of molecular and biochemical peculiarities discovered in these organisms that reflect their >1 billion years of separate evolution since the last common ancestor (Burki 2014; Cavalier-Smith 2010). Examples of their molecular novelty include a genome organised to facilitate polycistronic transcription of protein coding genes (Berriman, et al. 2005; El-Sayed, et al. 2005a; Ivens, et al. 2005; Martinez-Calvillo, et al. 2004; Martinez-Calvillo, et al. 2003), ubiquitous trans-splicing of a common short exon to the 5' end of all mRNAs (Sutton and Boothroyd 1986), transcription of some protein coding genes by RNA polymerase I in African trypanosomes (Gunzl, et al. 2003), RNA editing of mitochondrial transcripts (Aphasizhev and Aphasizheva 2014; Aphasizheva and Aphasizhev 2016), and possession of highly divergent chromosomal biology (Akiyoshi and Gull 2013; Ersfeld and Gull 1997), including unique epigenetic modifications (Siegel, et al. 2009) and kinetochore proteins (Akiyoshi and Gull 2014). All of these facets have combined to make trypanosomatid organisms of particular interest both in terms of both eukaryotic cell evolution and the host-parasite interactions.

Pathogenic kinetoplastid organisms are intensively studied to understand their biology and disease aetiology, but also for their immune evasion mechanisms that enable persistent infections.

Molecular interactions involved in evasion of host defences occur at the kinetoplastid cell surface and as a consequence the composition is specialised for each niche in the host. Thus, African

trypanosomes are extracellular and proliferate in blood and tissues fluids, and the cell surface is dominated by a single variant surface glycoprotein (VSG) packed to close to maximum possible density (Schwede, et al. 2015). The long-term survival of an infection is dependent on population level antigenic variation of the VSG, underpinned by a genomic repertoire of >1000 VSG coding sequences, comprising around 10% of all genes in the nuclear genome (Horn 2014; Schwede and Carrington 2010).

In contrast, both *T. cruzi* and *Leishmania* species proliferate inside host cells, a range of cell types for *T. cruzi*, and macrophages in the case of *Leishmania*. In *T. cruzi*, 6% of nuclear-encoded genes comprise diverse families of mucins (MUC I, MUC II and SMUG) that encode O-glycosylated proteins expressed in various combinations in the different developmental forms (Buscaglia, et al. 2006; Urban, et al. 2011). The initial O-glycosylation added during export to the cell surface is further modified at the cell surfaces by trans-sialidases that transfer sialic acid from host proteins to mucin acceptors, the sialylated oligosaccharides are then capped by a terminal α 1,3-galactose. The trans-sialidase gene family has expanded and diverged to a range of functions, many forms having lost catalytic activity (Nardy, et al. 2016). There are two other large gene families encoding cell surface proteins, amastins and MASPs, that are less well characterised (De Pablos and Osuna 2012; dos Santos, et al. 2012; Jackson 2010). In *Leishmania* sp., the cell surface is dominated by 'major surface protease' (MSP, also known as GP63 and leishmanolysin) surrounded by a sea of lipophosphoglycan and glycosylinositol phospholipids (Franco, et al. 2012; Yao, et al. 2003). MSP is a zinc metalloprotease encoded by a gene family that again has diverged to include both catalytically active and inactive members. As illustrated by the separate evolution of intracellular proliferation in *T. cruzi* and *Leishmania*, it is likely that the ability to proliferate intracellularly has evolved many times in different kinetoplastids and thus also likely involved separate evolution of immune evasion strategies.

In addition to the well characterised kinetoplastids described above, there is a large group of mammalian infective trypanosomatid species that are extracellular during proliferation in the mammalian host. In the vast majority of cases these species do not cause overt disease unless

host immunity is compromised. In contrast to African trypanosomes which can infect most species of mammal, these species have a relatively narrow host range. One example is *Trypanosoma (Megatrypanum) theileri* which infects Bovinae (cattle, buffalo, yaks and some antelopes) and is prevalent in cattle throughout the world (Farrar and Klei 1990; Garcia, et al. 2011; Greco, et al. 2000; Lee, et al. 2010; Matthews, et al. 1979; Niak 1978; Rodrigues, et al. 2003; Villa, et al. 2008). Infection with *T. theileri* normally results in a low parasitaemia that has been shown experimentally to be sustained for at least 12 weeks (Mott, et al. 2011) and is almost certainly lifelong. This indicates that the parasite can persist well beyond the impact of a primary immune response, demonstrating that effective immune evasion must be in operation, although the mechanisms that have evolved for evasion are unknown. Since parasite numbers in infected livestock can rapidly increase in immunocompromised, ill or stressed animals, the parasitaemia in healthy animals is probably limited by the host immune system (Townsend and Duffus 1985) and symptoms of disease are infrequent (Doherty, et al. 1993; Seifi 1995). This matches observations for African trypanosomes, such as *T. brucei*, that produce a higher parasitaemia but one that is also limited by both self-imposed quorum sensing and by the host, probably by innate and adaptive immune factors. Surveys in the US and Europe, and most recently the UK, indicate that *T. theileri* is present in >80% cattle but at very low parasitaemias within the blood and tissues of infected animals (Farrar and Klei 1990; Matthews, et al. 1979; Mott, et al. 2011; Schlafer 1979). *T. theileri* is transmitted by tabanid flies, where it undergoes a developmental cycle (Bose and Heister 1993). Infection of cattle is most likely mediated by ingestion of infected flies and also through vertical transmission. Related trypanosomes infect a range of mammals: *T. melophagium* in sheep, spread by keds (Gibson, et al. 2010); *T. pestanai* in badgers (Peirce and Neal 1974) spread by fleas (Lizundia, et al. 2011); *T. nabiasi* in rabbits (Grewal 1957) and *T. cervi* in deer (Matthews, et al. 1977) spread by keds (Bose and Petersen 1991).

Recent work has developed *T. theileri* as a potential vaccine delivery vehicle able to express antigens in sustained infections in recipient cattle and effective immune responses have been successfully generated to an encoded antigen (Mott, et al. 2011). These studies demonstrated that gene expression in *T. theileri* is similar to other trypanosomatids with polycistronic transcription of

protein coding genes followed by *trans*-splicing, cleavage and polyadenylation acting to resolve monocistronic mRNAs. Moreover, protein trafficking mechanisms were cross-functional with N-terminal signal sequences and GPI-addition sequences from *T. brucei* correctly targeting proteins to the exterior surface of the cell in *T. theileri*. In contrast, several expression regulatory elements that occur within the intergenic regions of polycistronic transcription in *Trypanosoma brucei* did not enable effective gene expression in *T. theileri*, suggesting that regulatory elements may not be conserved.

Here, we have derived a detailed picture of the genome and transcriptome of *T. theileri* in its mammalian infectious stage and present an analysis of its gene expression profile and a map of its RNA processing sites. These data allow us to predict that *T. theileri* is capable of meiotic exchange, gene silencing via RNA interference and that it contains conserved machinery for density-dependent growth control or cellular quiescence. In particular, the transcriptome has allowed a comparison of two distinct trypanosome cell surfaces, *T. brucei* and *T. theileri*, that have each evolved to enable the maintenance of a long-term extracellular infection in cattle. Notably, this provides evidence that *T. theileri* exploits a novel immune evasion mechanism distinct from that of the well characterised African trypanosome paradigm.

Materials and Methods

Genome sequencing and assembly

Trypanosoma theileri was isolated from a primary cell culture derived from a cow from the north west of England. *T. theileri* were cultured *in vitro* in (Mott, et al. 2011) and genomic DNA was extracted using Qiagen DNeasy Blood and Tissue kit. Isolated DNA was sequenced using a five library Illumina approach at the Beijing Genomics institute (www.genomics.cn/en/). The number of reads, read length and insert size of each library are shown in Supp. Table 1. Prior to assembly, reads were subject to quality filtering using trimmomatic (Bolger, et al. 2014) to remove low quality bases and read-pairs as well as contaminating adaptor sequences. Sequences were searched for all common Illumina adaptors (the default option) and the settings used for read processing by

trimmomatic were 'LEADING:10 TRAILING:10 SLIDINGWINDOW:5:15 MINLEN:50'. The quality filtered paired-end reads were then subject to assembly using ALLPATHS-LG (Maccallum, et al. 2009) using the default program settings. The resulting assembly was subject to 32 rounds of assembly error correction and gap filling using Pilon (Walker, et al. 2014) using the “–fix all” option and setting the expected ploidy to diploid. All filtered 91bp paired-end reads were mapped to this assembly set using BWA-MEM (Li and Durbin 2010), and read-pairs that did not map to the assembly were isolated and assembled separately using SGA (Simpson and Durbin 2012) using the default parameters. Contigs produced using SGA whose length was greater than 1000bp were added into the original assembly and subject to iterative scaffolding using SSPACE (Boetzer, et al. 2011). This process of identifying unmapped reads, assembly of unmapped reads and scaffolding was repeated until no further contigs greater than 1000bp were produced. The final draft assembly contained 319 sequences with an N50 515kb and a total assembly length of 29.8Mb and an average coverage per assembled contig of ~105x. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession NBCO000000000. The version described in this paper is version NBCO01000000.

Prediction of gene models

The assembled genome of *T. theileri* was subject to gene model prediction using Augustus (Stanke, et al. 2006). In brief, an initial set of gene models was predicted using gene prediction parameters inferred by a training gene model parameter using the set of genes currently annotated in the *T. cruzi* genome. These gene model parameters were used to predict a training set of genes in the draft assembly of *T. theileri*. The training set of genes were then used for multiple iterations of prediction and training until prediction converged on a final set of gene models and no further genes could be detected using Augustus. Gene models were also predicted using GeneMarkES (Besemer and Borodovsky 2005). GeneMarkES gene models were only kept if they did not overlap with an existing Augustus derived gene model and were consistent the gene orientation of their direct neighbour genes. The final set of predicted coding sequences comprised 11312 open reading frames.

Identification of polyadenylation and spliced-leader acceptor sites

T. theileri was grown as above and total RNA was extracted using the Qiagen RNeasy kit. Isolated RNA was sequenced at the Beijing Genomics institute (www.genomics.cn/en/) using two different approaches, a conventional TrueSeq protocol and a protocol designed to enrich for 3' ends of mRNA messages (Fiebig, et al. 2015). The number of reads, read length, and insert size of each library are shown in Supp. Table 1. The raw reads were used to identify and quantify the polyadenylation and spliced-leader acceptor sites in the genome of *T. theileri* using the SLaP mapper (Fiebig, et al. 2014) using the default settings. Raw RNASeq reads are available from EBI ArrayExpress under the accession number E-MTAB-5327.

Quantification of mRNA abundance

The conventional TruSeq reads above were also used to quantify the mRNA abundance for each predicted gene in the *T. theileri* genome. Prior to quantification reads were subject to quality filtering using trimmomatic as described above. The quality filtered paired-end reads were then used to quantify the abundance of the predicted gene models using RSEM (Li and Dewey 2011) utilising the default program parameters. mRNA abundances in *T. brucei* were derived from EBI ArrayExpress E-MTAB-3335 and E-MTAB-5460.

Inference of orthologous gene groups

The annotated proteins for *Trypanosoma brucei* TREU927, *T. congolense* IL3000, *T. cruzi* CL Brenner Esmereldo-like, *T. rangeli* SC58, *T. grayi* ANR4, *T. vivax* Y486, *Leishmania tarentolae* ParrotTarII, *L. major* Friedlin, *L. infantum* JPCM5, *L. mexicana* MHOMGT2001U1103, *L. donovani* BPK282A1, *L. braziliensis* MHOMBR71973M2269 and *Crithidia fasciculata* CfCI were downloaded from TriTrypDb V8.0 (Aslett, et al. 2010). The predicted proteins from *Phytomonas EM1* and *Phytomonas Hart1* were obtained from (Porcel, et al. 2014). The predicted proteins from *Phytomonas serpens* were obtained from (Koreny, et al. 2012) and for *Trypanoplasma borrelli* from (Kelly, et al. 2017). Orthologous gene groups were inferred using OrthoFinder (Emms and Kelly 2015) using default parameters.

Phylogenetic inference.

Orthologous gene groups containing only single copy genes in all species were selected for inclusion in the phylogenetic analysis. The individual orthologous groups were aligned using MergeAlign (Collingridge and Kelly 2012) and edited to remove all gap-containing columns and columns containing fewer than two character states. MergeAlign was selected as it has previously been shown to increase topological congruity between individual gene trees in multi-gene phylogenomic analyses (Collingridge and Kelly 2012). This resulted in a set of 375 alignments containing ≥ 127 phylogenetically informative un-gapped aligned positions in all 19 species. 100 bootstrapped concatenated multi-gene multiple sequence alignments were then constructed from this subset using an equi-sampling strategy. Each bootstrap replicate sampled, at random with replacement, 127 un-gapped phylogenetically informative columns from each multiple sequence alignment. This equi-sampling strategy was performed so that each gene contributed equally to the final phylogenetic tree thus preventing longer genes from biasing the result. Thus each bootstrap replicate contained 47,625 phylogenetically informative, un-gapped aligned positions spanning all species. Trees were inferred from each resampled multi-gene alignment using RAxML (Stamatakis 2006) utilising the PROTGAMMAAUTO model of sequence evolution and using minimum evolution principle (with log corrected scores) implemented in FastME (Desper and Gascuel 2002). The complete dataset including all alignments and phylogenetic trees is available from the Zenodo research data repository at <https://doi.org/10.5281/zenodo.193020>.

Results

Genome sequence, structure and cultured bloodstream form transcriptome

The *Trypanosoma theileri* genome was assembled using end reads from five paired end or jumping DNA libraries with insert sizes ranging from 170bp to 5000bp (Supp. Table 1). The genome contained 319 contigs and scaffolds with an N50 length of 515 kbp, a total length of 29.8 Mbp, and an average coverage depth of 105-fold (Supp. Figure 1). Subsequent analysis of the genome revealed that tandem repeats of near-identical genes were often absent from the genome

sequence and exhaustive attempts to develop *in silico* methods to assemble these were not successful. These tandemly repeated gene families remain a problem for a complete analysis of any Euglenozoan genome as many tandem repeats are compressed or absent from sequences assembled by large scale shotgun approaches using short reads, an example being the compression of the tubulin gene loci in the *T. brucei* reference genome (compare (Ersfeld, et al. 1998) and (Hall, et al. 2003)). After assembly of the *T. theileri* genome, 91.4% of the raw genome reads that had passed quality control mapped back to the assembly indicating that the assembly was relatively complete with respect to the input data. Several assembled scaffolds were equivalent in length to chromosomes present in other kinetoplastids (Supp. Figure 1), this highlighting the presence of clusters of directionally orientated genes indicative of polycistronic transcription units.

The total number of predicted protein coding genes in the genome assembly was 11312. This complement of protein coding genes was compared with other kinetoplastids by inferring orthologous sequence groups (orthogroups) (Emms and Kelly 2015) (Supp Table 2). The result was a list of orthogroups each containing more than one gene from one or more species. The *T. theileri* proteins fell into 8398 orthogroups and 9 of these were unique to *T. theileri*, including 5 orthogroups with more than ten members. Transcriptome data derived from bloodstream forms of *T. theileri* growing in axenic culture (Supp. Table 1) were used to validate gene models, and transcripts originating unambiguously from 11033 of the predicted genes was confidently detected. These transcriptome data were also used to map 5' *trans*-splice sites that act as recipients for the *T. theileri* spliced leader RNA (Rodrigues, et al. 2010), the short exon *trans*-spliced onto each mRNA molecule in kinetoplastid organisms. In addition, a dedicated transcriptome library was used to predict 3' polyadenylation sites (Fiebig, et al. 2014; Fiebig, et al. 2015) (Supp. Table 1). Together, these data allowed mapping of 5' *trans*-splice sites to 8585 genes and polyadenylation sites to 9160 genes and defined the processing sites and the untranslated regions (UTRs), which were then further analysed (Supp. Figure 2). For *trans*-splicing, the sequences 15 bp upstream and 10 bp downstream of the AG dinucleotide defining the splice leader addition site were analysed and showed little conservation (Supp. Figure 2). The majority of genes (~9000) contained

one or two splice acceptor sites with the remaining ~2000 genes containing between 3 and 9 splice acceptor sites; the length of the 5'UTR was less than 200 nucleotides for the vast majority of mRNAs (Supp. Figure 2). As with other kinetoplastids, the number of polyadenylation sites for each predicted gene was generally larger, multiple sites being the norm and the length of the 3'UTR also showed a greater range. The length distribution of 5'UTR (0-500nt) and 3'UTRs (50-1500nt) (Supp. Figure 2) was similar to that measured in *T. brucei* (Kolev, et al. 2010).

The transcriptome data was used to quantify mRNA levels as transcripts per million transcripts (TPM, the number of transcripts for a given gene per million mRNA transcripts) and this allowed a comparison of expression levels of genes/orthogroups between *T. theileri* and *T. brucei*. This revealed a good correlation between individual mRNA abundances in bloodstream forms of *T. theileri* and *T. brucei* when 5591 single copy orthologous genes were compared (Figure 1 and Supp. Table 2). One unexpected finding was a better correlation between *T. theileri* mammalian bloodstream form (BSF) and *T. brucei* insect procyclic form (PCF) transcripts than found between *T. theileri* BSF and *T. brucei* BSF transcripts (Spearman's coefficient $\rho = 0.46$ and $\rho = 0.57$ respectively, Monte Carlo resampling probability that the difference is chance $p < 0.001$). There are many possible explanations for the similarity, one is cell volume, with both *T. theileri* BSF and *T. brucei* PCF being significantly larger than *T. brucei* BSF. Presumably cell volume to surface area ratio has an effect on the relative abundance of many hundreds or thousands of cytoplasmic and cytoskeletal mRNAs. Another is that *T. brucei* may exhibit a particularly extreme developmental adaptation as a bloodstream form, exaggerating differences from its insect-dwelling forms, a feature not present in *T. theileri*, or possibly even other African trypanosome species.

Transcripts with the most different abundances in *T. theileri* and *T. brucei* BSFs were identified by selecting genes and orthogroups with more than an eight-fold difference in TPMs between the two species. This revealed that several glycosomal enzymes involved in the fate of the end products of glycolysis were differentially abundant in *T. brucei* when compared to *T. theileri* (Table 1). In *T. theileri* mRNAs for putative glycosomal isoforms of pyruvate phosphate dikinase, phosphoenolpyruvate carboxykinase and malate dehydrogenase were both ten-fold more

abundant than in *T. brucei* BSF while mRNAs encoding glycerol-3-phosphate dehydrogenase, glycerol kinase and mitochondrial alternative oxidase were at least ten-fold less abundant than in *T. brucei* BSFs. In the proliferating bloodstream form of *T. brucei*, ATP is generated by glycolysis and NAD^+ is regenerated through a cycle that uses the reduction of dihydroxyacetonephosphate (DHAP) to glycerol-3-phosphate to remove NADH and the alternative oxidase to convert glycerol-3-phosphate back to DHAP independently of NAD(H). Three carbon sugars enter this cycle as DHAP from glycolysis and leave on the conversion of glycerol-3-phosphate to glycerol generating ATP, and with glycerol being secreted. In contrast, in *T. theileri* mRNA expression levels are consistent with the conversion of pyruvate to PEP to oxaloacetate that re-enters the glycosome where it is converted to succinate allowing recovery of NAD^+ , the majority of succinate could then be secreted (Figure 2). This model deduced from differential mRNA levels is consistent with the measured conversion of glucose to succinate in bloodstream form *T. theileri* (van Hellemond, et al. 2007). Other than this metabolic distinction, there were no obvious further gross differences when sets of mRNAs corresponding to individual GO terms were compared.

The phylogenetic position of *T. theileri*

The predicted protein coding genes from *T. theileri* and 18 other kinetoplastids were used to infer orthogroups. Lists of genes in each orthogroup for each species are provided in Supp. Table 2. A phylogenomic analysis of 375 ubiquitously conserved single copy genes was conducted to investigate the relationship of *T. theileri* to other kinetoplastids with available genome sequences (Figure 3). This showed that *T. theileri* is positioned at the base of a monophyletic group containing the African crocodilian trypanosome *T. grayi* (Kelly, et al. 2014) as well as *T. cruzi* (El-Sayed, et al. 2005a) and *T. rangeli* (Stoco, et al. 2014). Thus, *T. theileri* is more closely related to these species than it is to African trypanosomes. This multi-gene view of the phylogenetic relationship between trypanosome species is compatible with previous studies that used fewer gene sequences for phylogenetic inference (Hamilton, et al. 2008; Hamilton, et al. 2004).

Interaction with the mammalian host and predicted cell surface

Three features characterise the abundant cell surface proteins on kinetoplastids: (i) a significant part of the genome is dedicated to the encoding genes, (ii) the proteins and mRNAs are very abundant, for example VSG is >10% of total cell protein, and (iii) the majority are attached to the plasma membrane via a GPI-anchor (the major exception being amastins in *Leishmania*). These three criteria were applied to the *T. theileri* genome and transcriptome datasets to identify genes encoding abundant cell surface proteins (Table 2). Homologues of genes present on the surface of other non-African trypanosomes and *Leishmania* sp. were present in *T. theileri*, including a large gene family encoding 216 distinct MSP-related proteins in orthogroup 2 (Supp. Table 2, OG00000002) and 36 genes encoding proteins with similarity to the trans-sialidase family (Supp. Table 2, OG00000005). In contrast to *T. cruzi*, there were relatively small numbers of amastin (Supp. Table 2, OG00000019) and MASP (Supp. Table 2, OG0007431) genes, and *T. brucei* VSG-like genes (Supp. Table 2, OG00000000) were not found. However, there were four large orthogroups, provisionally named *T. theileri* putative surface protein (TTPSP) 1 to 4, encoding proteins not present, or not readily detected, in other trypanosomatid species: 556 genes in OG00000004 encoding the TTPSP1 family, 30 in OG00000013 encoding TTPSP2, 145 in OG00000031 encoding TTPSP3, and 61 in OG0000108 encoding TTPSP4. Together, the MSP and four novel orthogroups contained 1008 genes, representing ~9% of the genes predicted to be present in the genome. All four novel orthogroups encoded polypeptides with conserved putative N-terminal signal and C-terminal GPI-addition sequences, represented by the logos in Figure 4a. All four also contained highly conserved residues close to the mature C-terminus that tended to be rich in serine residues but otherwise the mature polypeptides were divergent. The N-terminal signal and C-terminal GPI-anchor addition sequences from the four orthogroups all contained two or more cysteine residues and had some sequence similarity with each other. The TTPSP1 proteins were characterised by a conserved motif, including a run of threonine residues, close to the putative C-terminal GPI-anchor addition site which may be the substrate for O-glycosylation to produce mucin-like proteins, but otherwise had little conservation (Figure 4a, Supp. Figure 3a). The TTPSP2 proteins were also more conserved towards the C-terminus but without any obvious features (Figure 4a, Supp. Figure 3b). The TTPSP3 and 4 proteins had some regions with sequence conservation (Figure 4a, Supp. Figs. 3c and 3d) that may reflect conserved secondary

structure features. The four TTPSP families are novel and it remains to be determined whether they are simply structural or have additional activities. For example, although the TTPSP gene families had no overt similarity to VSGs, when the predicted structure of some members of TTPSP3 were analysed, for example Tth.6.1050, they yielded a high confidence (>95%) structural prediction with similarity to the *T. congolense* haptoglobin-haemoglobin receptor (HpHbR) (PDB 4E40) (Higgins, et al. 2013) (Supp. Figure 4). The HpHbR is an elongated three helical bundle and it has been proposed that a protein with this structure was the common ancestor of many African trypanosome cell surface proteins including HpHbR and GARP, an abundant protein in the surface coat of the tsetse forms of *T. congolense* (Higgins, et al. 2013). Transcription of the vast majority of the genes encoding MSP and the four TTPSP gene families was detected in RNAseq data. However, this RNAseq data was obtained from a population of cells in culture and thus it remains unknown whether all genes are expressed in all cells or whether individual cells express distinct cohorts of these genes.

The composition of the predicted cell surface (of the population) was estimated using transcript abundance as a proxy for protein abundance. To allow an approximate comparison between species, mRNA TPM values were normalised against alpha tubulin mRNA. For example, in bloodstream form *T. brucei* expressing VSG121, the VSG mRNA has a relative abundance to α -tubulin (Relative Abundance to α -Tubulin; RAT) of 9.9 as the TPM value for VSG mRNA is 9.9 times greater than α -tubulin mRNA. In *T. theileri*, five gene families encoding proteins with predicted GPI-anchors were expressed with transcript abundances close to or greater than α -tubulin (Table 2). The most abundant was MSP, RAT 5.32, followed by the four novel gene families: TTPSP4, RAT 1.73; TTPSP3, RAT 1.70; TTPSP2, RAT 1.44; and TTPSP1, RAT 1.17. In combination, these transcripts are 11.4 times more abundant than the α -tubulin mRNA, similar to the abundance of VSG mRNA in *T. brucei*. With the caveat that mRNA rather than protein has been measured, it can be predicted that the cell surface of *T. theileri*, irrespective of the precise cohort of genes that are expressed, is a densely packed coat of GPI-anchored proteins dominated by members of these five families.

The potential to modify cell surface molecules with carbohydrate was represented by the abundance of mRNAs for the addition of galactose with 60 genes encoding putative UDP-galactose/UDP-N-acetylglucosamine transferases; (OG000015, RAT 0.15) and two genes encoding galactokinases (OG000061, 2 genes, RAT 0.10) (Table 2). Both are predicted to be involved in mucin biosynthesis but also possibly in the modification of other surface glycans. A family of 36 genes encoding trans-sialidase homologues (OG0000005, RAT 0.12) contained potentially active members and some that are probably cell surface localised as they contain potential GPI-anchor addition sequences. The active site of trans-sialidases is characterised by a GRW motif and, of the 36 sequences, 30 spanned this motif and 9 of this 30 contained the motif G(R/K)W, whilst the majority of the others have a precise deletion of these three residues indicating that they are no longer enzymatically active (Supp. Figure 5).

In MSP, the presence of a conserved HExxHxxGF motif that binds the Zn^{++} ion in the active site was used as a measure of probable proteolytic activity. Of the 216 MSP sequences derived from the *T. theileri* genome, 158 spanned this motif, and of these 126 contained an intact motif (Supp. Figure 6). This observation suggests that the majority of MSP genes encode a potentially active protease, and thus proteolysis may have a role in immune evasion in *T. theileri*.

The vast majority of genes encoding putative cell surface proteins occurred in tandem arrays, some >100 kbp, that were distributed throughout the genome assembly (Supp. Figure 1). MSP, TTPSP3 and TTPSP4 genes were present in arrays containing just members of the same orthogroup. In contrast, TTPSP1 and TTPSP2 genes were interspersed in the same arrays (Figure 4b). Though it is clear that they represent a significant proportion of the genome, the full extent of the tandem arrays cannot be determined because of difficulties in assembling contigs containing large numbers of closely related genes using short reads.

***T. theileri* encodes genes for utilisation of exogenous sucrose**

A further gene family encoding invertase (glycosyl hydrolase/beta fructofuranosidase; OG0000080, 20 genes) (Table 2) was identified that is presumably involved in acquiring hexose from sucrose

in the gut of the insect vector. Although after mating female Tabanids become haematophagous they are still commonly found on flowers and feed on nectar (Kniepert 1980). While these invertases are related to the soluble extracellular sucrases of *Leishmania* (Gontijo, et al. 1996) (Supp Figure 7), all the copies in the *T. theileri* genome are predicted to be anchored to the cell surface by a GPI-anchor. In *Leishmania*, sucrases are expressed during the promastigote stage (Lyda, et al. 2015) and are thought to be important for proliferation of the parasite in the gut of their host insects (Blum and Oppendoes 1994), which also feed on plant juices. The predicted tethering of these sucrase enzymes to the *T. theileri* cell surface, via GPI-anchors, may enable the development of a locally higher glucose concentration at the cell surface where it can be actively transported into the cell. The presence of these sucrase genes in *T. theileri*, and in *Leishmania* sp. (Figure 3) indicates that the invertase gene family must have been lost from other lineages during the diversification of trypanosomes, possibly subsequent to adapting to a vector that does not feed on plant sugars.

Conserved features in trypanosomes

Analysis of the *T. theileri* genome highlights a number of predicted surface proteins conserved amongst all *Trypanosoma* species. This conserved cell surface group includes: 46 genes encoding adenylyl cyclase homologous to *T. brucei* ESAG4 (OG0000008) (Alexandre, et al. 1996) and 17 genes encoding ESAG3 homologues (OG0000025) (Pays, et al. 1989), both gene families contain many fragmentary copies as in *T. brucei*. Among other molecules universally conserved, *T. theileri* contains a highly conserved homologue of PSSA-2 (OG0000908) (Jackson, et al. 1993), as well as the aforementioned MSP, transialidase, and the amastin families. No apparent orthologues of the *T. brucei* transferrin receptor or the haptoglobin-haemoglobin receptor were detected in the *T. theileri* genome. However, it was found that mRNA encoding a putative haem transporter, LHR1 (OG0006128), was >10 fold higher in *T. theileri* than in *T. brucei* BSFs (Table 2) and this may provide haem and iron.

Host Cell Invasion in *T. cruzi*

A characteristic distinction between different kinetoplastid parasites is the capacity of some species to proliferate as an intracellular form in a variety of mammalian cell types. While African trypanosomes are exclusively extracellular in all stages of their life cycle, and *T. theileri* and *T. grayi* have only been detected in the extracellular milieu, *T. cruzi* has evolved the ability to invade mammalian cells replicating as intracellular amastigote forms. The molecules that have been described as linked to intracellular invasion by *T. cruzi* include trans-sialidase family members, Trypomastigote small surface antigen (TSSA), DGF, TcSMP, and members of the SAP protein family (Epting, et al. 2010; Maeda, et al. 2012; Martins, et al. 2015). A comparison of genomes of *T. theileri*, *T. grayi*, *T. rangeli* and *T. cruzi* was carried out to explore the conservation of cell surface protein families and, by identifying those unique to *T. cruzi*, potentially necessary for host cell invasion (Supplementary Table 3). Orthogroups containing *T. cruzi* mucins, SAP and TSSA are restricted to *T. cruzi*, whereas homologues of the other molecules were present in *T. theileri*. One possible interpretation is that SAP and TSSA evolved to facilitate a unique step cell invasion but there is evidence that many other proteins are involved, including those conserved in non-invading species. Finally, it is possible that one or more of the TTPSP gene families shares a common origin with *T. cruzi* mucins but that they were not placed in the same orthogroup.

Life cycle

A recent genome wide survey of genes linked to density-dependent developmental events in the *T. brucei* life cycle identified a cohort of genes involved in quorum sensing and/or quiescence (Mony, et al. 2014). It is not known whether *T. theileri* has a mechanism to restrict uncontrolled proliferation, although cell cycle arrest in preparation for developmental progression is a common feature of kinetoplastid life cycles. Analysis of the *T. theileri* genome identified many of the signalling components identified in *T. brucei* that are involved in developmental arrest (Mony and Matthews 2015) including genes including AMPK (Saldivia, et al. 2016) and YAK protein kinases (OG0005762 and OG0001009 respectively), protein phosphatase 1 (OG0000045) and an orthologue of the TbRBP7 predicted RNA binding protein (OG0000232), over-expression of which drives premature differentiation to stumpy forms in the mammalian bloodstream. The presence of these genes could indicate a form a density dependent growth restraint in *T. theileri* operates, or

the capacity for cellular quiescence, which is a ubiquitous but diverse feature of eukaryotic life (O'Farrell 2011). Whether this is linked to developmental progression however is unknown. Nonetheless, molecules implicated in other differentiation events in the trypanosome life cycle were detected in the *T. theileri* genome including the tyrosine phosphatase TbPTP1 (OG0000462) (Szoor, et al. 2006), the glycosomal serine threonine phosphatase TbPIP39 (OG0000450) (Szoor, et al. 2010), and the RNA regulators of development, TbZFP1, 2 and 3 (Walrad, et al. 2009). Other signalling molecules linked to developmental events included NrKA/B (Domingo-Sananes, et al. 2015; Gale, et al. 1994; Gale and Parsons 1993) (OG0000380), which promotes development upon entry into the tsetse fly in *T. brucei*, and RDK1 and RDK2 (Jones, et al. 2014) kinases (OG0000116 and OG0000231 respectively) which inhibit differentiation to procyclic forms; MAPK2, associated with proliferation of differentiated procyclic forms in *T. brucei* was also detected (OG0005019) (Muller, et al. 2002). Further evidence for a developmental cycle within the arthropod vector is the presence of genes uniquely expressed during, and necessary for, meiosis. Thus, all four of the genes, SPO11 (OG0003147), MND1 (OG0003699), HOP1 (OG0002840) and DMC1 (OG004601) (Peacock, et al. 2011) are present in the *T. theileri* genome. With respect to the morphological transformations that accompany the developmental cycle, the trypanosome flagellum attachment zone associated cytoskeletal protein GM6 (OG0001193) (Hayes, et al. 2014) was present in *T. theileri*. Each of these analyses provides evidence for a functional developmental cycle in *T. theileri*, distinct from simple mechanical transmission. The predicted underlying molecular controls are also similar to those identified in *T. brucei*.

Signalling and the Cell cycle

To explore conservation of the cell cycle machinery we first analysed the presence of molecules required for normal cell growth or viability previously characterised in *T. brucei*. Orthologues of the 43 protein kinases that have been found by RNAi analysis in *T. brucei* to be required for normal growth or viability are all present in the *T. theileri* genome (Alsford, et al. 2011; Jones, et al. 2014). Of genes with a predicted protein kinase function in *T. theileri*, four have no orthologue in *T. brucei*, with Tth.16.1150 (OG0008376) and Tth.10.1250 (OG0002653) having an orthologue in *T. cruzi* but not *Leishmania* and *Phytomonas*. Tth.23.1270 (OG0005077) is a further protein kinase absent in

the African trypanosomes but present in *T. grayi* as well as *T. theileri*, whilst Tth.37.2060 (OG0002304) has a small open reading frame with other related and longer members of the encoded protein family detected in the genomes of *T. grayi*, *T. congolense* and *T. vivax*. Progression through the cell cycle is regulated by cyclin dependent kinases, and all the cyclin and kinase components identified in *T. brucei* were present in the *T. theileri* genome.

RNA interference

The RNAi machinery shows variable presence amongst different kinetoplastid species. To be operational a core set of 5 proteins are required for RNAi, these being AGO1 (OG0005445), DCL1 (OG0006584), DCL2 (OG0007047), RIF4 (OG0007336) and RIF5 (OG0006482). All of these genes were detected in the *T. theileri* genome, indicating an intact RNAi based gene silencing machinery.

Discussion

Nearly every vertebrate that has been specifically investigated has been found to harbour one or more trypanosomatid species and most proliferate in the blood and sometimes tissue spaces of their hosts. The best-characterised are *T. brucei* and *T. cruzi* as both can cause human disease. However, these are not typical of the genus as the majority of species have evolved neither VSG-based antigenic variation, as found in African trypanosomes such as *T. brucei*, nor the ability to proliferate inside a host cell like *T. cruzi*. Most are also not overtly pathogenic.

T. theileri represents a more typical trypanosome where overt disease symptoms are an unusual sequelae of natural infection. It has a narrow host range, infecting cattle and other Bovinae, and can maintain a low level infection for months and probably for the lifetime of the host. *T. theileri* provides an informative contrast to *T. brucei*: although both have evolved the ability to infect cattle in parallel, *T. theileri* is a Stercorarian trypanosome, whereas *T. brucei* is a Salivarian, these clades having separated before the emergence of Bovinae from other mammals (Stevens and Gibson 1999). Unlike *T. theileri*, *T. brucei* is characterised by a large host range including cattle, and it is

possible that cattle have only been available as a host since the introduction of domesticated cattle to sub-Saharan Africa in the last ten thousand years, although there are several other Bovinae species amongst the indigenous fauna. Nonetheless, both trypanosome species can maintain persistent extracellular infections in cattle and, at least in an African context, would often coexist simultaneously in coinfections. Here, a genome and transcriptome of *T. theileri* has been determined to provide an insight into the norm with respect to the trypanosomatid lifestyle and as an informative comparator with more pathogenic and better characterised trypanosomes. The main findings are:

- (i) the structure of the genome and the features of the transcriptome is typical for a kinetoplastid.
- (ii) A phylogenetic analysis using sequences of 375 polypeptides universally conserved in trypanosomes confirms earlier work placing *T. theileri* with the Stercorarian trypanosomes, closer to *T. cruzi*, *T. rangeli* and *T. grayi*, than to Salivarian trypanosomes such as *T. brucei*.
- (iii) The core cellular machinery is conserved and comparison of relative transcript abundance with *T. brucei* identified few major differences, the main exception being in the fate of glycolytic end products with possible secretion of succinate in *T. theileri* compared to secretion of glycerol in *T. brucei*, consistent with previous measurements (van Hellemond, et al. 2007).
- (iv) An analysis of transcript abundance allowed a prediction that the *T. theileri* cell surface is dominated by five polypeptide families, MSP and four novel genes families, TTPSP1 to TTPSP4. Transcripts encoding GPI-anchored trans-sialidases and galactose transferases able to modify N- and O-linked oligosaccharides were abundant.
- (v) A complete life cycle was indicated by the presence of a full complement of genes involved in cell cycle arrest, developmental transitions and meiosis. It can be predicted that *T. theileri* is competent for RNAi.
- (vi) A comparison of the genomes of Stercorarian trypanosomes to find putative orthologues of genes implicated in cell invasion in *T. cruzi* identified TSSA and SAP genes being unique to *T. cruzi*. These two gene families may have evolved specifically for host cell invasion.

The structure of the *T. theileri* genome was typical for a kinetoplastid (El-Sayed, et al. 2005b). Genes are arranged in closely spaced tandem arrays consistent with polycistronic transcription

from occasional start sites and linked processing to monocistronic mRNAs through *trans*-splicing to add a short capped exon at the 5' end (the spliced leader) and linked cleavage and polyadenylation. This model was supported by the identification of RNA processing sites for spliced leader addition and polyadenylation with the resolution of the transcriptome enabling detection of intergenic sequences as well as mature mRNAs. Further experimental support comes from the efficacy of transgenic dicistronic transcription units in *T. theileri* (Mott, et al. 2011).

Analysis of the cellular core machinery encoded in the genome revealed the expected conservation of fundamental cellular processes. The function of the identified genes was analysed by assigning the encoded polypeptides to orthogroups (Emms and Kelly 2015). An orthogroup is the set of genes descended from a single copy gene in the last common ancestor of the species being analysed, and thus orthogroups, like orthologs, are a natural unit for comparison between species. To add depth to this analysis, we compared the expression levels of mRNAs from both single copy orthologous genes and shared orthogroups in *T. brucei* and *T. theileri*. This approach compared two species that proliferate at a similar rate in culture and have independently evolved to grow in the same host. With the proviso that mRNA levels are only a proxy for the relative importance of pathways or pathway components, and translational control will certainly be important, the analysis suggested general similarity. However, the main identifiable difference was that the end products of glycolysis differ between the two species with production and possible secretion of succinate in *T. theileri* as opposed to the secretion of glycerol in *T. brucei* bloodstream forms (van Hellemond, et al. 2007).

Whilst the core cellular functions or adaptations to different environments are expected to be similar between the trypanosome species, the most significant differences will lie in their interactions with their hosts including regulation of parasitaemia and immune evasion. These interactions are mediated by components present on trypanosome cell surfaces that are usually characterised by densely packed coats of GPI-anchored proteins and oligo- and/or polysaccharides in some species. In the case of African trypanosomes, the coat is dominated by a single variant surface glycoprotein (VSG) packed to close to maximum possible physical density

(Manna, et al. 2014; Schwede, et al. 2015). Outside African trypanosomes, the best characterised kinetoplast cell surfaces are from *T. cruzi*, the causal agent of human American trypanosomiasis, which proliferates inside a range of host cells, and various *Leishmania* species, the causal agents of a range of diseases and characterised by an intracellular proliferative cycle in macrophages. In *T. cruzi*, 6% of genes encode diverse families of mucins (MUC I, MUC II and SMUG) that encode O-glycosylated proteins expressed in various developmental forms (Buscaglia, et al. 2006). The initial O-glycosylation added during biosynthesis is further modified at the cell surfaces by trans-sialidases that transfer sialic acid from host proteins to mucin acceptors. The modified oligosaccharides are capped by a terminal α 1,3-galactose. In addition, there are two other large gene families encoding cell surface proteins, amastins and MASPs, that are less well characterised. In *Leishmania*, the cell surface is dominated by MSP (Yao 2010) surrounded by a sea of lipophosphoglycan and glycosylinositol phospholipids (Ilg 2000; Medina-Acosta, et al. 1993). To gain insight into the *T. theileri* cell surface, we analysed the abundant cell surface proteins predicted by mRNA levels. This revealed the presence of a large family of MSP proteins and four other novel families, TTPSP1 to TTPSP4, that are very diverse in sequence except at their N- and C-termini. The MSP family contained potentially proteolytically active members as well as inactive members. Hence, the cell surface composition of MSP and the four novel protein families represents a distinct architecture to the monotonous simplicity of the VSG in *T. brucei*. Although MSP is well characterised, structural and functional characteristics of the proteins encoded by the other four families were not readily apparent. However, TTPSP1, the most numerous gene family, encoded proteins with runs of ~10 threonine residues in a location proximal to the putative GPI-anchor and thus plasma membrane. This suggests an extracellular domain that is likely to be O-glycosylated and subsequently modified by cell surface trans-sialidases. The combination of a densely packed coat containing sialic acid modified mucins and proteolytically active MSPs is a potent cocktail and could well enable the parasite to escape immune recognition. Potential MSP substrates include immunoglobulins and complement components attached to the trypanosome surface. The importance of these surface families is reflected in their overall representation in the genome, where the number of genes encoding predicted major surface proteins totals over one thousand. However, this represents a minimum estimate since difficulties

in the assembly of the genome sequence for tandemly arrayed gene families means that the full extent of these tandem gene arrays is unclear. Nonetheless, it is interesting that the number of genes encoding abundant putative cell surface proteins represent approximately 10% of the coding capacity of the genome, similar to the overall contribution of VSG genes to the genome of *T. brucei*. The combined expression level of the mRNAs of the surface proteins predicted in *T. theileri* also approximates to the contribution of the VSG mRNA to the overall transcript abundance in a *T. brucei* cell. However, it is important to note that it is not clear whether the diverse gene families are uniformly expressed in individual cells or different members of the population express distinct representatives of the families. Nonetheless, the genomic organisation of these gene families coupled with the detection of expression of the majority of members indicates that most can be expressed and that the cell surface is complex.

Although the vast majority are detected as expressed, the genomic arrangement of TTPSP genes into tandem arrays is reminiscent of the arrays of silent VSG gene arrays in *T. brucei*. The TTPSP and VSG gene families also share other similar features. For example, both the *T. theileri* gene families and the *T. brucei* VSGs have moderately well conserved N-terminal signal and C-terminal GPI-anchor addition sequences but are highly divergent for nearly all of the mature polypeptide sequence. Moreover, some members of TTPSP2 have confident predictions for a three helical bundle structure present in several African trypanosome cell surface proteins and proposed to be the origin of the VSG N-terminal domain fold (Higgins, et al. 2013). The presence of proteins with this fold in both African trypanosomes and *T. theileri* indicates that three helical bundle cell surface proteins were present in the common ancestor of *T. theileri* and *T. brucei*. With these observations in mind, it is possible to speculate that the key step in the evolution of VSG-based immune evasion that evolved in African trypanosomes was the evolution of RNA polymerase I transcription of a proto-VSG from a pre-existing diverse gene family. Such events have been suggested before (Jackson and Barry 2012; Manna, et al. 2014) but this the first evidence that three helical bundle proteins pre-dated VSGs.

VSG expression in one of multiple telomeric sites (Horn 2014) and the co-evolution of diverse expression associated gene families might have enabled the large host range exhibited by *T. brucei*. In contrast, the more restricted host range of *T. theileri* probably results from one or more essential molecular interactions with host macromolecules that are conserved in Bovinae but have diverged in other mammal families. With the current state of knowledge, it is not possible to predict what these interactions are from genome and transcriptome sequence data alone.

Interrogation of the *T. theileri* genome identified genes encoding orthologues of several proteins with experimentally determined functions in the developmental cycle of *T. brucei*. Molecules linked to quorum sensing and cellular quiescence were identified, as were molecules associated with life in the arthropod vector, including those associated with sexual exchange. The implication is that *T. theileri* is normally competent for a full developmental cycle in an arthropod vector, most frequently tabanid flies although ticks have also been proposed to be competent for transmission (Morzaria, et al. 1986). Moreover, the conservation of molecules linked to the morphological events and developmentally regulated alteration in cell surface proteins was evidenced by the presence of the GM6 and PSSA2, both of which are present throughout the genus. Interestingly, although the genome and transcriptome of the *T. theileri* is distinct from *T. brucei*, a regulator of stumpy form development in *T. brucei* in the bloodstream, RBP7, is also present and syntenic in *T. theileri*, with two closely related isoforms present in tandem, as in *T. brucei*. The master regulator of development through the tsetse fly, RBP6, is also present and syntenic in *T. theileri*. Taken together this suggests that conserved regulatory mechanisms underly the developmental cycle of trypanosome parasites, which nonetheless differ in their developmental morphotypes.

Finally, one feature of kinetoplastid parasites that has been either retained or lost in the different representatives is the presence of a functional RNAi machinery. This is defined by the presence of five core proteins, AGO, DCL1, DCL2, RIF 4 and RIF5. All of these components are detectable within the *T. theileri* genome, as in *T. grayi*, but unlike *T. cruzi* and *T. rangeli* where it is absent. The presence of the RNAi machinery, the genome sequence and the amenability to culture means

that many of the predictions and hypotheses above will be testable in the future by laboratory manipulation.

In conclusion, this analysis has provided evidence for an alternative means to evade the mammalian immune systems as a replicative extracellular trypanosome distinct from the VSG-based antigenic variation characterised in African trypanosomes. Particularly, these successful parasites occupy the same host, the same niche, and commonly at the same time as the pathogenic African trypanosomes. The respective evolutionary trade-offs between immune evasion mechanisms, level of parasitaemia in the host and transmission efficiency using distinct arthropod vectors in each of these trypanosome species represents an interesting area of comparative further study.

Acknowledgements

This work was supported by the Biotechnology and Biological Sciences Research Council [BB/F00057X/1 to KM, BB/L02442X/1 to KM]; the Wellcome Trust grants [103740 to KM, 095831 to KM, 085256 to MCa]; an EU Horizon 2020 award [637765 to SK]. SK is a Royal Society University Research Fellow; MCI and KT are recipients of an RCUK-CONFAP research partnership award [BB/M029239/1]. Next-generation sequencing and library construction was delivered via the BBSRC National Capability in Genomics [BB/J010375/1] at the Earlham Institute (EI, formerly The Genome Analysis Centre, Norwich), by members of the Platforms and Pipelines Group. Bioinformatics support was funded via BBSRC Institute Strategic Programme grant [BB/J004669/1] to the Earlham Institute.

Table 1

Comparison of mRNA expression levels, expressed as transcripts per million transcripts (TPM), for enzymes involved in glycolysis and the resolution of products of glycolysis in *T. theileri* bloodstream forms (Tt BSF) and *T. brucei* bloodstream forms (Tb BSF). In addition, a comparison of the haem transporter LHR1 is shown at the bottom of the Table. Increased abundance in *T. theileri* relative to *T. brucei* is shown in blue and decreased in red.

Enzyme	Tb gene		Tt gene	Tb BSF	Tt BSF	Tb BSF / Tt BSF	Tt BSF / Tb BSF
Glycerol-3-phosphate dehydrogenase	Tb927.1.1130	cytosolic	Tth.14.2510	29	17	1.71	0.59
	Tb927.8.3530	glycosome	Tth.25.1530	1022	33	30.97	0.03
	Tb927.11.7380	mitochondrion	Tth.16.2900	100	31	3.23	0.31
Glycerol kinase	Tb927.9.12550	glycosome	Tth.4.3700	872	48	18.17	0.06
	Tb927.9.12570						
	Tb927.9.12590						
	Tb927.9.12610						
	Tb927.9.12630						
Alternative Oxidase	Tb927.10.7090	mitochondrion	Tth.8.1700	244	21	11.62	0.09
	Tb927.10.9760						
Pyruvate phosphate dikinase	Tb927.11.6280	glycosome	Tth.16.1420	41	680	0.06	16.59
Phosphoenolpyruvate carboxykinase	Tb927.2.4210	glycosome	Tth.23.1750	116	2421	0.05	20.87
Malate dehydrogenase	Tb927.10.2560	mitochondrion	Tth.1.1170	72	434	0.17	6.03
	Tb927.10.15410	glycosome	Tth.3.2160	81	1394	0.06	17.21
	Tb927.11.11250	cytosolic	Tth.24.1250	176	826	0.21	4.69
Fumarate hydratase	Tb927.3.4500	glycosome	Tth.9.3160	28	164	0.17	5.86
	Tb927.11.5050	mitochondrion	Tth.43.1640	70	43	1.63	0.61
Fumarate reductase	Tb927.5.940	glycosome	Tth.75.1060	17	251	0.07	14.76
			Tth.88.1120				
	Tb927.5.930		Tth.75.1070	28	220	0.13	7.86
	Tb927.10.3650		Tth.75.1080				
Succinate dehydrogenase	Tb927.8.6580	mitochondrion	Tth.42.1190	43	56	0.77	1.30
Succinate dehydrogenase	Tb927.8.6580	mitochondrion	Tth.15.2280	57	190	0.30	3.33
LHR1 haem uptake protein	Tb927.8.6010			47	532	0.09	11.35

Table 2

T. theileri: putative cell surface proteins and modifying enzymes. Expression levels, in TPM, relative to alpha tubulin mRNA, and in rank of abundance are shown, as is the presence of orthologues in other species.

Orthogroup		Gene copy number in orthogroup	Orthogroup mRNA level (sum TPM)	mRNA abundance relative to alpha tubulin	mRNA rank	<i>T. theileri</i>	<i>T. grayi</i>	<i>T. cruzi</i>	<i>T. rangeli</i>	<i>T. brucei</i>	<i>Leishmania</i>
<i>T. theileri</i>											
OG0000130	alpha tubulin	?	18047	1.00	13	✓	✓	✓	✓	✓	✓
conserved cell surface proteins											
OG0000002	MSP	216	95986	5.32	1	✓	✓	✓	✓	✓	✓
OG0000019	amastin	2	498	0.03	244	✓	✓	✓	✓	✓	✓
OG0000908	PSSA-2	5	200	0.01	664	✓	✓	✓	✓	✓	○
OG0007431	MASP	1	491	0.03	252	✓	✓	✓	✓	○	○
unique cell surface proteins											
OG0000004	TTPSP1	556	21178	1.17	6	✓	○	○	○	○	○
OG0000013	TTPSP2	304	26071	1.44	5	✓	○	○	○	○	○
OG0000031	TTPSP3	145	30737	1.70	4	✓	○	○	○	○	○
OG0000108	TTPSP4	61	31188	1.73	3	✓	○	○	○	○	○
oligosaccharide modifying enzymes											
OG0000015	UDP-Gal or UDP-GlcNAc-dependent glycosyltransferase	60	2631	0.15	39	✓	✓	✓	✓	✓	?
OG0000061	galactokinase	2	1789	0.10	71	✓	✓	✓	✓	✓	✓
OG0000080	glycosyl hydrolase / beta fructofuranosidase sucrase	20	512	0.03	239	✓	✓	✓	✓	✓	✓
OG0000005	trans-sialidase	36	2227	0.12	54	✓	✓	✓	✓	✓	○
<i>T. brucei</i> for comparison											
OG0000130	alpha tubulin	~20		1	5	✓	✓	✓	✓	✓	✓
OG0000000	VSG	1		9.93	1	○	○	○	○	✓	○

References

- Akiyoshi B, Gull K 2014. Discovery of unconventional kinetochores in kinetoplastids. *Cell* 156: 1247-1258. doi: 10.1016/j.cell.2014.01.049
- Akiyoshi B, Gull K 2013. Evolutionary cell biology of chromosome segregation: insights from trypanosomes. *Open Biol* 3: 130023. doi: 10.1098/rsob.130023
- Alexandre S, et al. 1996. Families of adenylate cyclase genes in *Trypanosoma brucei*. *Mol Biochem Parasitol* 77: 173-182.
- Alsford S, et al. 2011. High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome. *Genome Res* 21: 915-924. doi: 10.1101/gr.115089.110
- Aphasizhev R, Aphasizheva I 2014. Mitochondrial RNA editing in trypanosomes: small RNAs in control. *Biochimie* 100: 125-131. doi: 10.1016/j.biochi.2014.01.003
- Aphasizheva I, Aphasizhev R 2016. U-Insertion/Deletion mRNA-Editing Holoenzyme: Definition in Sight. *Trends Parasitol* 32: 144-156. doi: 10.1016/j.pt.2015.10.004
- Aslett M, et al. 2010. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* 38: D457-462. doi: 10.1093/nar/gkp851
- Azizi MH, Bahadori M, Dabiri S, Shamsi Meymandi S, Azizi F 2016. A History of Leishmaniasis in Iran from 19th Century Onward. *Arch Iran Med* 19: 153-162. doi: 0161902/AIM.0016
- Berriman M, et al. 2005. The genome of the African trypanosome *Trypanosoma brucei*. *Science* 309: 416-422. doi: 10.1126/science.1112642
- Besemer J, Borodovsky M 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* 33: W451-454. doi: 10.1093/nar/gki487
- Blum JJ, Opperdoes FR 1994. Secretion of sucrase by *Leishmania donovani*. *J Eukaryot Microbiol* 41: 228-231.
- Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27: 578-579. doi: 10.1093/bioinformatics/btq683
- Bolger AM, Lohse M, Usadel B 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114-2120. doi: 10.1093/bioinformatics/btu170
- Bose R, Heister NC 1993. Development of *Trypanosoma (M.) theileri* in tabanids. *J Eukaryot Microbiol* 40: 788-792.

- Bose R, Petersen K 1991. *Lipoptena cervi* (Diptera), a potential vector of *Megatrypanum* trypanosomes of deer (Cervidae). *Parasitol Res* 77: 723-725.
- Burki F 2014. The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb Perspect Biol* 6: a016147. doi: 10.1101/cshperspect.a016147
- Buscaglia CA, Campo VA, Frasch AC, Di Noia JM 2006. *Trypanosoma cruzi* surface mucins: host-dependent coat diversity. *Nat Rev Microbiol* 4: 229-236. doi: 10.1038/nrmicro1351
- Cavalier-Smith T 2010. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biol Lett* 6: 342-345. doi: 10.1098/rsbl.2009.0948
- Collingridge PW, Kelly S 2012. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC Bioinformatics* 13: 117. doi: 10.1186/1471-2105-13-117
- Cox FE 2016. The Golden Age of parasitology-1875-1925: the Scottish contributions. *Parasitology*: 1-15. doi: 10.1017/S0031182016001566
- Crooks GE, Hon G, Chandonia JM, Brenner SE 2004. WebLogo: a sequence logo generator. *Genome Res* 14: 1188-1190. doi: 10.1101/gr.849004
- De Pablos LM, Osuna A 2012. Multigene families in *Trypanosoma cruzi* and their role in infectivity. *Infect Immun* 80: 2258-2264. doi: 10.1128/IAI.06225-11
- Desper R, Gascuel O 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comput Biol* 9: 687-705. doi: 10.1089/106652702761034136
- Doherty ML, et al. 1993. Clinical disease associated with *Trypanosoma theileri* infection in a calf in Ireland. *Vet Rec* 132: 653-656.
- Domingo-Sananes MR, Szoor B, Ferguson MA, Urbaniak MD, Matthews KR 2015. Molecular control of irreversible bistability during trypanosome developmental commitment. *J Cell Biol* 211: 455-468. doi: 10.1083/jcb.201506114
- dos Santos SL, et al. 2012. The MASP family of *Trypanosoma cruzi*: changes in gene expression and antigenic profile during the acute phase of experimental infection. *PLoS Negl Trop Dis* 6: e1779. doi: 10.1371/journal.pntd.0001779
- El-Sayed NM, et al. 2005a. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* 309: 409-415. doi: 10.1126/science.1112631

- El-Sayed NM, et al. 2005b. Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309: 404-409. doi: 10.1126/science.1112181
- Emms DM, Kelly S 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 16: 157. doi: 10.1186/s13059-015-0721-2
- Epting CL, Coates BM, Engman DM 2010. Molecular mechanisms of host cell invasion by *Trypanosoma cruzi*. *Exp Parasitol* 126: 283-291. doi: 10.1016/j.exppara.2010.06.023
- Ersfeld K, Asbeck K, Gull K 1998. Direct visualisation of individual gene organisation in *Trypanosoma brucei* by high-resolution in situ hybridisation. *Chromosoma* 107: 237-240.
- Ersfeld K, Gull K 1997. Partitioning of large and minichromosomes in *Trypanosoma brucei*. *Science* 276: 611-614.
- Farrar RG, Klei TR 1990. Prevalence of *Trypanosoma theileri* in Louisiana cattle. *J Parasitol* 76: 734-736.
- Fiebig M, Gluenz E, Carrington M, Kelly S 2014. SLaP mapper: a webserver for identifying and quantifying spliced-leader addition and polyadenylation site usage in kinetoplastid genomes. *Mol Biochem Parasitol* 196: 71-74. doi: 10.1016/j.molbiopara.2014.07.012
- Fiebig M, Kelly S, Gluenz E 2015. Comparative Life Cycle Transcriptomics Revises *Leishmania mexicana* Genome Annotation and Links a Chromosome Duplication with Parasitism of Vertebrates. *PLoS Pathog* 11: e1005186. doi: 10.1371/journal.ppat.1005186
- Franco LH, Beverley SM, Zamboni DS 2012. Innate immune activation and subversion of Mammalian functions by *leishmania* lipophosphoglycan. *J Parasitol Res* 2012: 165126. doi: 10.1155/2012/165126
- Gale M, Jr., Carter V, Parsons M 1994. Translational control mediates the developmental regulation of the *Trypanosoma brucei* Nrk protein kinase. *J Biol Chem* 269: 31659-31665.
- Gale M, Jr., Parsons M 1993. A *Trypanosoma brucei* gene family encoding protein kinases with catalytic domains structurally related to Nek1 and NIMA. *Mol Biochem Parasitol* 59: 111-121.
- Garcia HA, et al. 2011. High genetic diversity in field isolates of *Trypanosoma theileri* assessed by analysis of cathepsin L-like sequences disclosed multiple and new genotypes infecting cattle in Thailand. *Vet Parasitol* 180: 363-367. doi: 10.1016/j.vetpar.2011.03.017

- Gibson W, Pilkington JG, Pemberton JM 2010. *Trypanosoma melophagium* from the sheep ked *Melophagus ovinus* on the island of St Kilda. *Parasitology* 137: 1799-1804. doi: 10.1017/S0031182010000752
- Gontijo NF, Melo MN, Riani EB, Almeida-Silva S, Mares-Guia ML 1996. Glycosidases in *Leishmania* and their importance for *Leishmania* in phlebotomine sandflies with special reference to purification and characterization of a sucrase. *Exp Parasitol* 83: 117-124. doi: 10.1006/expr.1996.0055
- Greco A, Loria GR, Dara S, Luckins T, Sparagano O 2000. First isolation of *Trypanosoma theileri* in Sicilian cattle. *Vet Res Commun* 24: 471-475.
- Grewal MS 1957. The life cycle of the British rabbit trypanosome, *Trypanosoma nabiasi* Railliet, 1895. *Parasitology* 47: 100-118.
- Gunzl A, et al. 2003. RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *Eukaryot Cell* 2: 542-551.
- Hall N, et al. 2003. The DNA sequence of chromosome I of an African trypanosome: gene content, chromosome organisation, recombination and polymorphism. *Nucleic Acids Res* 31: 4864-4873.
- Hamilton PB, Adams ER, Malele, II, Gibson WC 2008. A novel, high-throughput technique for species identification reveals a new species of tsetse-transmitted trypanosome related to the *Trypanosoma brucei* subgenus, *Trypanozoon*. *Infect Genet Evol* 8: 26-33. doi: 10.1016/j.meegid.2007.09.003
- Hamilton PB, Stevens JR, Gaunt MW, Gidley J, Gibson WC 2004. Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA. *Int J Parasitol* 34: 1393-1404. doi: 10.1016/j.ijpara.2004.08.011
- Hayes P, et al. 2014. Modulation of a cytoskeletal calpain-like protein induces major transitions in trypanosome morphology. *J Cell Biol* 206: 377-384. doi: 10.1083/jcb.201312067
- Higgins MK, et al. 2013. Structure of the trypanosome haptoglobin-hemoglobin receptor and implications for nutrient uptake and innate immunity. *Proc Natl Acad Sci U S A* 110: 1905-1910. doi: 10.1073/pnas.1214943110
- Horn D 2014. Antigenic variation in African trypanosomes. *Mol Biochem Parasitol* 195: 123-129. doi: 10.1016/j.molbiopara.2014.05.001

- Ilg T 2000. Proteophosphoglycans of Leishmania. *Parasitol Today* 16: 489-497.
- Ivens AC, et al. 2005. The genome of the kinetoplastid parasite, *Leishmania major*. *Science* 309: 436-442. doi: 10.1126/science.1112680
- Jackson AP 2010. The evolution of amastin surface glycoproteins in trypanosomatid parasites. *Mol Biol Evol* 27: 33-45. doi: 10.1093/molbev/msp214
- Jackson AP, Barry JD. 2012. The Evolution of Antigenic Variation in African Trypanosomes. In: Sibley LD, Howlett BJ, Heitmann J, editors. *Evolution of Virulence in Eukaryotic Microbes.*: John Wiley & Sons. p. 324-337.
- Jackson DG, Smith DK, Luo C, Elliott JF 1993. Cloning of a novel surface antigen from the insect stages of *Trypanosoma brucei* by expression in COS cells. *J Biol Chem* 268: 1894-1900.
- Jones NG, et al. 2014. Regulators of *Trypanosoma brucei* cell cycle progression and differentiation identified using a kinome-wide RNAi screen. *PLoS Pathog* 10: e1003886. doi: 10.1371/journal.ppat.1003886
- Kelly S, Carrington M, Dóro E, Forlenza M, G. W 2017. Transcriptome sequence of the bloodstream form of *Trypanoplasma borreli*, a haematozoic parasite of fish transmitted by leeches. *Genome Announcements*: in press.
- Kelly S, Ivens A, Manna PT, Gibson W, Field MC 2014. A draft genome for the African crocodilian trypanosome *Trypanosoma grayi*. *Sci Data* 1: 140024. doi: 10.1038/sdata.2014.24
- Kniepert FW 1980. Blood feeding and nectar-feeding in adult Tabanidae (Diptera). *Oecologia* 46: 125-129.
- Kolev NG, et al. 2010. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. *PLoS Pathog* 6: e1001090. doi: 10.1371/journal.ppat.1001090
- Koreny L, et al. 2012. Aerobic kinetoplastid flagellate *Phytomonas* does not require heme for viability. *Proc Natl Acad Sci U S A* 109: 3808-3813. doi: 10.1073/pnas.1201089109
- Lee YF, et al. 2010. Isolation of *Trypanosoma (Megatrypanum) theileri* from dairy cattle in Taiwan. *J Vet Med Sci* 72: 417-424.
- Li B, Dewey CN 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12: 323. doi: 10.1186/1471-2105-12-323

- Li H, Durbin R 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26: 589-595. doi: 10.1093/bioinformatics/btp698
- Lizundia R, et al. 2011. Evidence for a role of the host-specific flea (*Paraceras melis*) in the transmission of *Trypanosoma (Megatrypanum) pestanai* to the European badger. *PLoS One* 6: e16977. doi: 10.1371/journal.pone.0016977
- Lyda TA, et al. 2015. A unique, highly conserved secretory invertase is differentially expressed by promastigote developmental forms of all species of the human pathogen, *Leishmania*. *Mol Cell Biochem* 404: 53-77. doi: 10.1007/s11010-015-2366-6
- Maccallum I, et al. 2009. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol* 10: R103. doi: 10.1186/gb-2009-10-10-r103
- Maeda FY, Cortez C, Yoshida N 2012. Cell signaling during *Trypanosoma cruzi* invasion. *Front Immunol* 3: 361. doi: 10.3389/fimmu.2012.00361
- Manna PT, Boehm C, Leung KF, Natesan SK, Field MC 2014. Life and times: synthesis, trafficking, and evolution of VSG. *Trends Parasitol* 30: 251-258. doi: 10.1016/j.pt.2014.03.004
- Martinez-Calvillo S, Nguyen D, Stuart K, Myler PJ 2004. Transcription initiation and termination on *Leishmania major* chromosome 3. *Eukaryot Cell* 3: 506-517.
- Martinez-Calvillo S, et al. 2003. Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. *Mol Cell* 11: 1291-1299.
- Martins NO, et al. 2015. Molecular Characterization of a Novel Family of *Trypanosoma cruzi* Surface Membrane Proteins (TcSMP) Involved in Mammalian Host Cell Invasion. *PLoS Negl Trop Dis* 9: e0004216. doi: 10.1371/journal.pntd.0004216
- Matthews DM, Kingston N, Maki L, Nelms G 1979. *Trypanosoma theileri* Laveran, 1902, in Wyoming cattle. *Am J Vet Res* 40: 623-629.
- Matthews MJ, Kingston N, Morton JK 1977. *Trypanosoma cervi* Kingston and Morton, 1975 from mule deer, *Odocoileus hemionus*, in Wyoming. *J Wildl Dis* 13: 33-39.
- Medina-Acosta E, Beverley SM, Russell DG 1993. Evolution and expression of the *Leishmania* surface proteinase (gp63) gene locus. *Infect Agents Dis* 2: 25-34.
- Mony BM, et al. 2014. Genome-wide dissection of the quorum sensing signalling pathway in *Trypanosoma brucei*. *Nature* 505: 681-685. doi: 10.1038/nature12864

- Mony BM, Matthews KR 2015. Assembling the components of the quorum sensing pathway in African trypanosomes. *Mol Microbiol* 96: 220-232. doi: 10.1111/mmi.12949
- Morzaria SP, Latif AA, Jongejan F, Walker AR 1986. Transmission of a *Trypanosoma* sp. to cattle by the tick *Hyalomma anatolicum anatolicum*. *Vet Parasitol* 19: 13-21.
- Mott GA, et al. 2011. Targeting cattle-borne zoonoses and cattle pathogens using a novel trypanosomatid-based delivery system. *PLoS Pathog* 7: e1002340. doi: 10.1371/journal.ppat.1002340
- Muller IB, Domenicali-Pfister D, Roditi I, Vassella E 2002. Stage-specific requirement of a mitogen-activated protein kinase by *Trypanosoma brucei*. *Mol Biol Cell* 13: 3787-3799. doi: 10.1091/mbc.E02-02-0093
- Nardy AF, Freire-de-Lima CG, Perez AR, Morrot A 2016. Role of *Trypanosoma cruzi* Transsialidase on the Escape from Host Immune Surveillance. *Front Microbiol* 7: 348. doi: 10.3389/fmicb.2016.00348
- Niak A 1978. The incidence of *Trypanosoma theileri* among cattle in Iran. *Trop Anim Health Prod* 10: 26-27.
- O'Farrell PH 2011. Quiescence: early evolutionary origins and universality do not imply uniformity. *Philos Trans R Soc Lond B Biol Sci* 366: 3498-3507. doi: 10.1098/rstb.2011.0079
- Pays E, et al. 1989. The genes and transcripts of an antigen gene expression site from *T. brucei*. *Cell* 57: 835-845.
- Peacock L, et al. 2011. Identification of the meiotic life cycle stage of *Trypanosoma brucei* in the tsetse fly. *Proc Natl Acad Sci U S A* 108: 3671-3676. doi: 10.1073/pnas.1019423108
- Peirce MA, Neal C 1974. *Trypanosoma (Megatrypanum) pestanai* in British badgers (*Meles meles*). *Int J Parasitol* 4: 439-440.
- Porcel BM, et al. 2014. The streamlined genome of *Phytomonas* spp. relative to human pathogenic kinetoplastids reveals a parasite tailored for plants. *PLoS Genet* 10: e1004007. doi: 10.1371/journal.pgen.1004007
- Rodrigues AC, et al. 2003. Brazilian isolates of *Trypanosoma (Megatrypanum) theileri*: diagnosis and differentiation of isolates from cattle and water buffalo based on biological characteristics and randomly amplified DNA sequences. *Vet Parasitol* 116: 185-207.

- Rodrigues AC, et al. 2010. Characterization of spliced leader genes of *Trypanosoma* (*Megatrypanum*) *theileri*: phylogeographical analysis of Brazilian isolates from cattle supports spatial clustering of genotypes and parity with ribosomal markers. *Parasitology* 137: 111-122. doi: 10.1017/S0031182009991053
- Saldivia M, Ceballos-Perez G, Bart JM, Navarro M 2016. The AMPKalpha1 Pathway Positively Regulates the Developmental Transition from Proliferation to Quiescence in *Trypanosoma brucei*. *Cell Rep* 17: 660-670. doi: 10.1016/j.celrep.2016.09.041
- Schlafer DH 1979. *Trypanosoma theileri*: a literature review and report of incidence in New York cattle. *Cornell Vet* 69: 411-425.
- Schwede A, Carrington M 2010. Bloodstream form Trypanosome plasma membrane proteins: antigenic variation and invariant antigens. *Parasitology* 137: 2029-2039. doi: 10.1017/S0031182009992034
- Schwede A, Macleod OJ, MacGregor P, Carrington M 2015. How Does the VSG Coat of Bloodstream Form African Trypanosomes Interact with External Proteins? *PLoS Pathog* 11: e1005259. doi: 10.1371/journal.ppat.1005259
- Seifi HA 1995. Clinical trypanosomosis due to *Trypanosoma theileri* in a cow in Iran. *Trop Anim Health Prod* 27: 93-94.
- Siegel TN, et al. 2009. Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. *Genes Dev* 23: 1063-1076. doi: 10.1101/gad.1790409
- Simpson JT, Durbin R 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* 22: 549-556. doi: 10.1101/gr.126953.111
- Stamatakis A 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688-2690. doi: 10.1093/bioinformatics/btl446
- Stanke M, Tzvetkova A, Morgenstern B 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* 7 Suppl 1: S11 11-18. doi: 10.1186/gb-2006-7-s1-s11
- Stevens JR, Gibson W 1999. The molecular evolution of trypanosomes. *Parasitol Today* 15: 432-437.

- Steverding D 2008. The history of African trypanosomiasis. *Parasit Vectors* 1: 3. doi: 10.1186/1756-3305-1-3
- Steverding D 2014. The history of Chagas disease. *Parasit Vectors* 7: 317. doi: 10.1186/1756-3305-7-317
- Stoco PH, et al. 2014. Genome of the avirulent human-infective trypanosome, *Trypanosoma rangeli*. *PLoS Negl Trop Dis* 8: e3176. doi: 10.1371/journal.pntd.0003176
- Sutton RE, Boothroyd JC 1986. Evidence for trans splicing in trypanosomes. *Cell* 47: 527-535.
- Szoor B, Ruberto I, Burchmore R, Matthews KR 2010. A novel phosphatase cascade regulates differentiation in *Trypanosoma brucei* via a glycosomal signaling pathway. *Genes Dev* 24: 1306-1316. doi: 10.1101/gad.570310
- Szoor B, Wilson J, McElhinney H, Taberner L, Matthews KR 2006. Protein tyrosine phosphatase TbPTP1: A molecular switch controlling life cycle differentiation in trypanosomes. *J Cell Biol* 175: 293-303. doi: 10.1083/jcb.200605090
- Townsend J, Duffus WP 1985. Antibody-dependent cellular cytotoxicity of *Trypanosoma theileri* mediated by purified bovine isotypes and subisotypes. *Parasite Immunol* 7: 179-189.
- Urban I, et al. 2011. Molecular diversity of the *Trypanosoma cruzi* TcSMUG family of mucin genes and proteins. *Biochem J* 438: 303-313. doi: 10.1042/BJ20110683
- van Hellemond JJ, et al. 2007. Energy metabolism of bloodstream form *Trypanosoma theileri*. *Eukaryot Cell* 6: 1693-1696. doi: 10.1128/EC.00130-07
- Villa A, et al. 2008. Presence of *Trypanosoma theileri* in Spanish Cattle. *Ann N Y Acad Sci* 1149: 352-354. doi: 10.1196/annals.1428.016
- Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9: e112963. doi: 10.1371/journal.pone.0112963
- Walrad P, Paterou A, Acosta-Serrano A, Matthews KR 2009. Differential trypanosome surface coat regulation by a CCCH protein that co-associates with procyclin mRNA cis-elements. *PLoS Pathog* 5: e1000317. doi: 10.1371/journal.ppat.1000317
- Yao C 2010. Major surface protease of trypanosomatids: one size fits all? *Infect Immun* 78: 22-31. doi: 10.1128/IAI.00776-09

Yao C, Donelson JE, Wilson ME 2003. The major surface protease (MSP or GP63) of *Leishmania* sp. Biosynthesis, regulation of expression, and function. Mol Biochem Parasitol 132: 1-16.

Figure Legends

Figure 1

Comparison of mRNA levels, for 5591 common single copy genes, between: *T. theileri* cultured bloodstream forms (BSF) and (a) *T. brucei* cultured BSFs and (b) *T. brucei* cultured procyclic (insect) forms (PCF). (c) A comparison between the two *T. brucei* life cycle stages. mRNA abundance is shown as Transcripts per Million Transcripts (TPM).

Figure 2

The measurements from Table 2 shown on the glycolytic pathway in trypanosomes. The blue spots indicate enzymes encoded by mRNAs that have >10-fold higher expression (measured as TPM) in *T. theileri* BSFs than in *T. brucei* BSFs and the red spots < 10-fold lower expression in *T. theileri*.

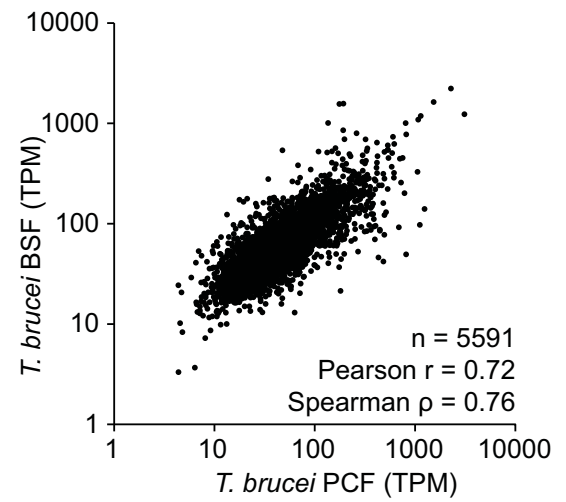
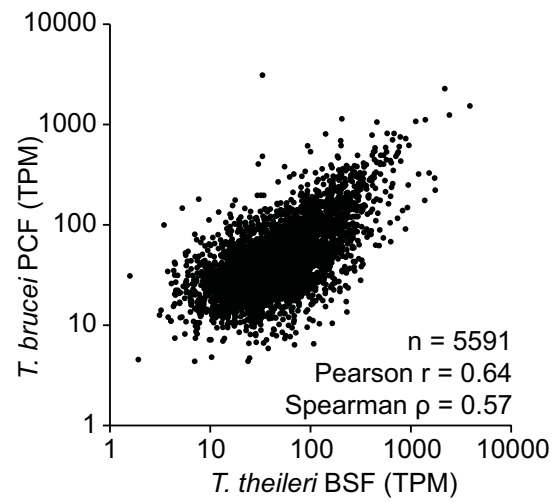
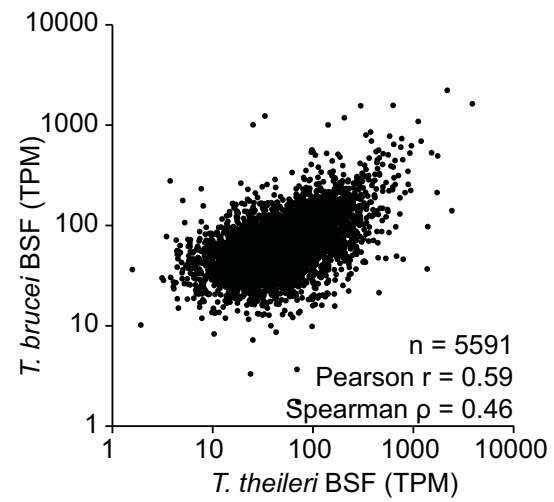
Figure 3

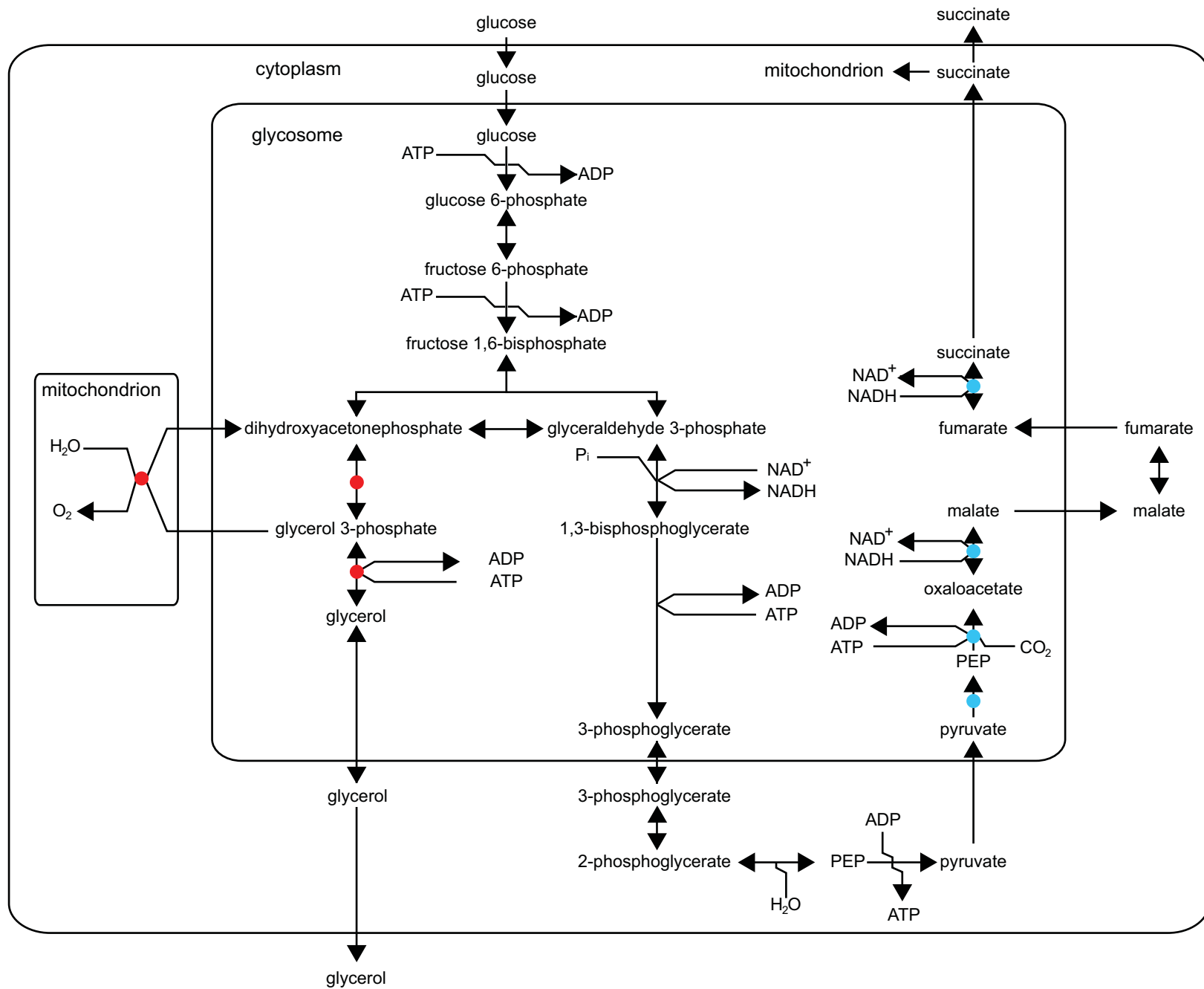
Phylogenetic relationship between Kinetoplastida with substantial genome/transcriptome sequence availability. The relationship is based on a equal-sampled concatenated alignment of 375 ubiquitously conserved single copy genes (see Methods). Bootstrap replicate support values (percentages) are shown at internal nodes. Scale bar indicates the number of substitutions per site.

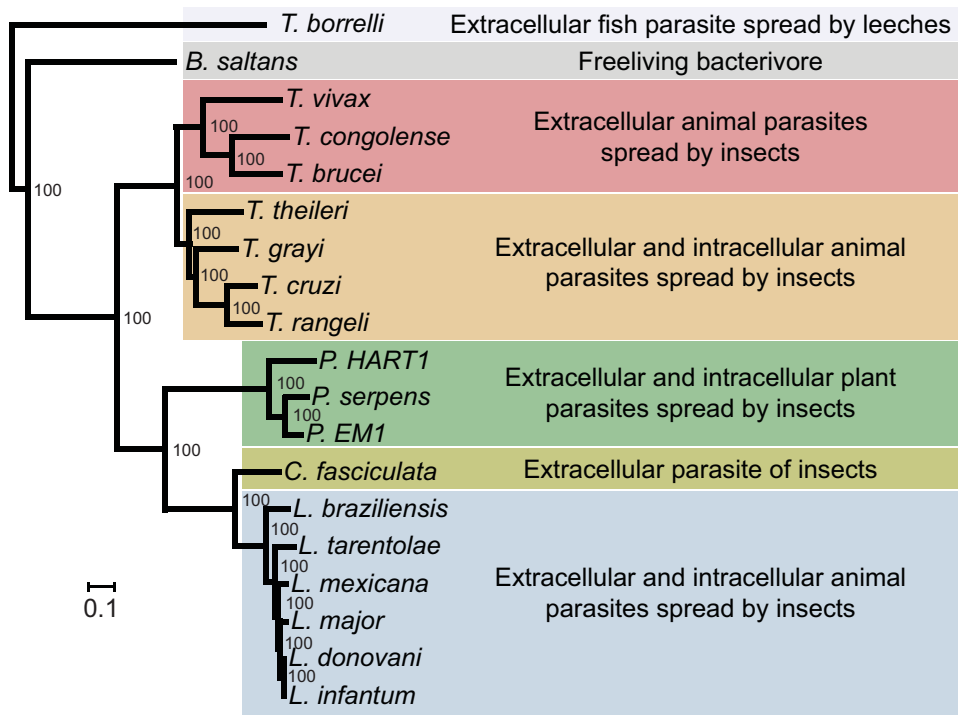
Figure 4

(a) Four novel gene families in *T. theileri* encode polypeptides with conserved signal sequences and C-termini. The comparisons were made using Logos (Crooks, et al. 2004) after alignment of either the N- or C-termini.

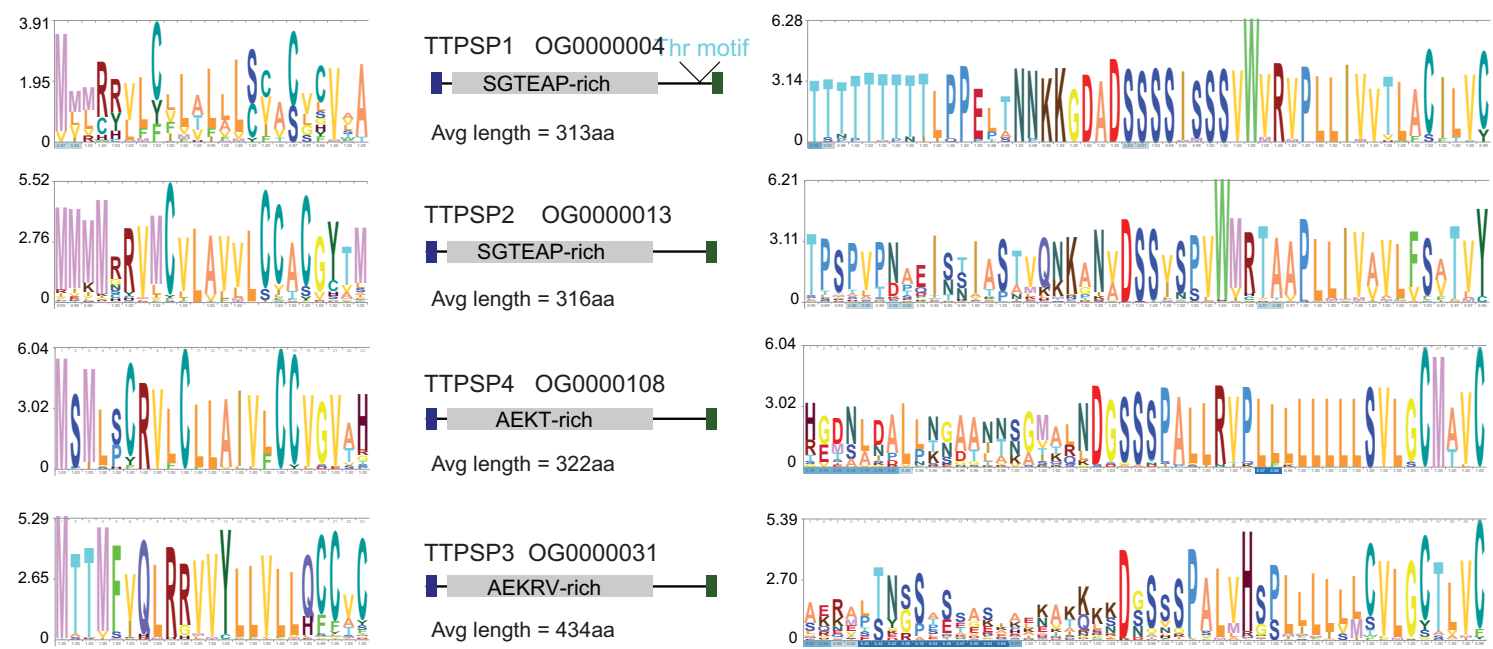
(b) Genomic structure of a tandem array of interspersed genes encoding members of OG000004 and OG000013, two orthogroups encoding abundant putative cell surface proteins.







a



b

